VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

Medicare fraud detection

Project Report

BIG DATA FRAMEWORKS

(CSE3120)

School of Computer Science and Engineering (SCOPE)

VIT Chennai

Winter 2022-2023

Slot: G1

Course Faculty: Prof Mansoor Hussain D

Submitted by

RISHI RANJAN – 20MIA1004

JEEVESH NANDAN UPADHAYA – 20MIA1116

ACKNOWLEDGEMENT

We would like to express our special thanks and gratitude to our teacher **Prof Mansoor Hussain D**, faculty of school of Computer Science and Engineering who gave us the golden opportunity to do this wonderful project titled **medical fraud detection**, and also for providing us with proper guidance and suggestions to develop this project. This helped us in research and we came across many new things. Secondly, we would also like to thank our friends and group mates who helped us a lot in finalizing this project within the limited time frame. We would also like to thank our family for their consistent support throughout this time.

Our exploratory analysis on Medicare fraud detection involves building and assessing three learners on each dataset. Based on the Area under the Receiver Operating Characteristic (ROC) Curve performance metric, our results show that the Combined dataset with the Logistic Regression (LR) learner yielded the best overall score at 0.816, closely followed by the Part B dataset with LR at 0.805. Overall, the Combined and Part B datasets produced the best fraud detection performance with no statistical difference between these datasets, over all the learners. Therefore, based on our results and the assumption that there is no way to know within which part of Medicare a physician will commit fraud, we suggest using the Combined dataset for detecting fraudulent behaviour when a physician has submitted payments through any or all Medicare parts evaluated in our study.

CONTENTS

ABSTRACT

Access to affordable health care is a national issue that affects most people in the United States. Medicare is a Federal Government health care program that helps people over 65 and individuals get affordable health insurance.

with some kinds of handicaps. The Medicare system is plagued by fraud, waste, and abuse that jeopardizes the health and well-being of beneficiaries while costing taxpayers billions of dollars annually. Work done in the past has shown that publicly available Medicare claims data can be used to build machine learning models that can automatically spot fraud. However, problems with big data that aren't balanced by class hurt performance.

INTRODUCTION

Healthcare is important to many people in the United States (U.S.), but the high costs of health-related services mean that many people can't get the care they need. Because of this, the U.S. government has set up and paid for programmes, such as Medicare, that help people who need medical care but don't have enough money to pay for it. There are a number of problems with healthcare and medical insurance systems, such as a growing population or bad actors (such as fraudulent or potentially fraudulent doctors or providers), which reduces the amount of money set aside for these programmes. The number of people 65 and older in the U.S. has grown by 28% from 2004 to 2015, while the number of people under 65 has grown by only 6.5%. One reason for this is that healthcare has gotten better, which has helped people live longer. The U.S. healthcare spending has gone up, with an annualised growth rate of 4.0% (adjusted for inflation) between 1995 and 2015. This is because the population has grown, especially among the elderly, and medical technology has improved. Most likely, spending will keep going up, making it even more important to have an efficient and cost-effective healthcare system. Fraud, waste, and abuse are big problems in healthcare, and even though people are trying to stop them, they aren't making a big difference in the amount of money they cost. In this study, we look at fraud, and when we say "fraud" in this paper, we also mean "waste" and "abuse." The Federal Bureau of Investigation (FBI) says that 3–10% of healthcare costs are due to fraud, which costs between $19 billion and $65 billion each year. Medicare accounts for 20% of all healthcare spending in the U.S. If effective fraud detection methods were used, Medicare alone could recover between $3.8 billion and $13 billion in costs.

The majority of the time, auditors and investigators in the healthcare industry will manually sift through a large number of records in order to look for potentially fraudulent or suspicious actions. This is the primary method for detecting fraud. This laborious procedure, which involves sifting through vast volumes of data,

through, which can be time-consuming and relatively ineffective in comparison to more automated methods such as data mining and machine learning when it

comes to detecting fraud. Because technology advancements now make it possible to store large amounts of data, for example, in electronic health records (EHR), the volume of information that is used in the healthcare industry is continuing to grow. This is making "Big Data" an increasingly important concept. The ability to perform data mining and machine learning on Big Data is increasing along with the advancement and increased usage of technology. This has the potential to enhance the current condition of healthcare as well as medical insurance programmes so that patients can receive quality medical care. The Canters for Medicare Services (CMS) participated in this initiative by publishing "Big Data" Medicare datasets to aid in the identification of fraud, waste, and abuse that occurs inside Medicare. According to a statement issued by CMS, "those individuals who are set on misusing Federal health care programmes can cost taxpayers billions of dollars while putting beneficiaries' health and welfare in jeopardy." As Medicare continues to provide services to an increasing number of people, the impact of these losses and hazards will become more severe. On the website for the Centres for Medicare Services (CMS), there are numerous datasets that users can access.

The five Vs of Big Data are Volume, Velocity, Variety, Veracity, and Value. Volume refers to large volumes of data, Velocity refers to the rapid rate at which new data is generated/collected, and Variety refers to the level of complexity of the data (for example, incorporating

Veracity symbolises the genuineness of the data, and Value indicates how good the quality of the data is in relation to the expected results.

CMS's databases display many of these Big Data characteristics. These databases are Huge Volume because they contain annual claim records for all physicians in the United States who submit to Medicare. CMS provides data for the prior year every year, increasing the Big Volume of available data. The datasets each have roughly 30 variables, ranging from provider demographics and procedure kinds to payment amounts and the number of services performed, making them Big Variety.

Furthermore, because it integrates three significant (but distinct) Medicare data sources, the Combined dataset used in our analysis automatically contains Big Variety data. We believe that these datasets are credible, valid, and representative of all known Medicare provider claims since CMS is a

government programme with clear quality controls and thorough documentation for each dataset, showing Great Veracity. According to research undertaken by our research department and others, this data can be utilised to detect fraudulent behaviour, giving it a high value. Furthermore, because it contains the largest known repository of real-world fraudulent medical providers in the United States, the LEIE dataset could be termed Big Value.

This research makes two contributions. Initially, we present in-depth discussions on Medicare Big Data processing as well as exploratory experiments and analysis to demonstrate the best learners and datasets for detecting Medicare provider claims fraud. Data imputation, deciding which variables (dataset characteristics) to preserve, aggregating the data from the procedure-level to the provider-level to match the level of the LEIE dataset for fraud label mapping, and constructing the Combined dataset are our unique data processing steps.

It should be noted that the fraud labels are used to assess fraud by leveraging previous exclusion information as well as Medicare payments made to currently excluded providers. Second, the resulting processed datasets are referred to as Big Data, and as a result, for our fraud detection experiments, we utilize these datasets.

Use Spark on top of a Hadoop YARN cluster capable of handling these large dataset sizes. The four Medicare datasets were trained and verified using fivefold cross-validation for our experiments, and the process was repeated ten times. We build the Random Forest (RF), Gradient Tree Boosting (GTB), and Logistic Regression (LR) models from the Apache Spark 2.3.0 Machine Learning Library, and utilize the Area under the ROC Curve (AUC) statistic to assess fraud detection performance. We chose these learners because they are widely used and provide reasonably acceptable performance for our exploratory research of fraud detection performance in Medicare using Big Data. We evaluate statistical significance utilizing the Analysis Of Variance (ANOVA) and Tukey's Honest Significant Difference (HSD) tests to add robustness to the results. According to our findings, the Combined dataset with LR produced the greatest overall AUC of 0.816, followed by the Part B dataset with LR at 0.805. Furthermore, the Part B dataset produced the greatest results for GBT and RF, both with a 0.796 AUC.

The DMEPOS dataset produced the weakest fraud detection results, with RF having the lowest overall AUC of 0.708. The results for the Combined dataset using LR show that it outperforms any individual Medicare dataset; consequently, the whole is greater than the sum of its parts in this scenario. This is not the case for RF or GBT, where Part B has the highest average AUC. Yet, there was no statistical difference between the Combined dataset and the Part B dataset results. As a result, the high fraud detection rates, along with our premise that Medicare fraud can occur in any or all components of Medicare, show the possibility for using the Combined dataset to successfully detect provider claims fraud across learners. To conclude, the following are the paper's unique contributions:

•• Explicitly describing Medicare Part B, Part D, and DMEPOS data processing, as well as real-world fraud label mapping.

•• Merging the three Medicare large datasets into a single Consolidated dataset to exhibit high fraud detection performance that takes into consideration all of Medicare's important components.

•• Investigating the performance of fraud detection and learner behaviour in each of the four large datasets.

The remainder of the paper is structured as follows. The "Related Works" section discusses similar works, with a focus on works that use several CMS branches of Medicare. The "Datasets" section goes over the various Medicare datasets that are used, how the data is processed, and the fraud label mapping approach. The "Methods" section describes the methods that were employed, such as the learners, performance metric, and hypothesis testing. The section "Results and comments" explains the outcomes of our investigation. Lastly, in the "Conclusion" section, we wrap up and discuss future work.

LITERATURE SURVEY

When detecting fraudulent behaviour, the overwhelming majority of these studies rely solely on Medicare Part B data, neglecting to account for other Medicare programmes. Anywhere within the healthcare system where funds are anywhere money is being exchanged, there is an opportunity for a bad actor to manipulate the process and siphon funds, affecting the efficiency and effectiveness of the Medicare healthcare process.

It is possible for a bad actor to manipulate the process and syphon off funds, thereby influencing the efficiency and effectiveness of the Medicare healthcare process.

There is limited prior information regarding where (in the Medicare system) a physician will commit fraud; therefore, selecting a particular Medicare component may overlook fraud committed elsewhere. This study focuses on the processing and categorization of each Medicare dataset, as well as the performance of fraud detection. Therefore, we restrict our discussion in this section to the limited literature that attempts to identify fraudulent behaviour using multiple CMS datasets. As of this research, only two works fit into this category.

Utilize the Part B (2012–2014), Part D (2013), and LEIE datasets in this instance.

They do not specify how they preprocess the data or combine Part B and Part D, but they accept attributes from both Part B and Part D datasets and treat drugs and HCPCS codes identically. They identified 12,153 fraudulent physicians using the National Provider Identifier (NPI) and their proprietary algorithm for correlating identities.

They chose not to differentiate between LEIE exclusion rules/codes and instead used every physician listed. It is unclear whether the authors accounted for waivers, exclusion start dates, or exclusion duration in their process for mapping fraud labels. These particulars are essential for reducing redundant and overlapping exclusion labels and evaluating the efficacy of fraud detection systems. Due to this lack of clarity in the exclusion labelling methodology, the results of their study cannot be reproduced reliably and are difficult to compare to other studies. They devised a method for identifying fraudulent behaviour by applying network algorithms to graphs to determine the fraud risk. Due to the

extremely unbalanced nature of the data, the authors utilised a 50:50 class distribution, retaining 12,000 excluded providers and selecting 12,000 non-excluded providers at random. They presented several groups of algorithms and based their fraud detection outcomes on the real-world fraudulent physicians discovered in the LEIE dataset. Using nominal values such as medication prescriptions and medical procedures, one set of algorithms, which they refer to as Behaviour–Vector similarity, establishes behavioural similarity between real-world fraudulent and non-fraudulent physicians. A second set of algorithms comprise their risk propagation, which employs geospatial co-location (such as practise location) to estimate the risk propagation from fraudulent healthcare providers. An ablation analysis revealed that the majority of this predictive accuracy was due to characteristics that assess risk propagation via geospatial collocation.

DATASET

This section describes the CMS datasets we employ.

In addition, the data processing methodology used to generate each dataset, including processing, fraud label mapping between the Medicare datasets and the LEIE, and one hot processing, must be disclosed.

We discuss encoding for categorical variables. Each dataset's information is derived from administrative claims data collected by CMS for Medicare beneficiaries enrolled in the Fee-For-Service program. Note that this information does not include claims submitted through the Medicare Advantage program. We presume the Medicare data is already cleansed and accurate given that CMS records all claim information after payments are made. Note that the NPI is not utilized during data mining, but rather for aggregation and identification. In addition, we added a year variable to each dataset, which is also used for aggregation and identification.

Medicare dataset descriptions

Phase B

The Part B data set contains claims information for each procedure performed by a physician during a given year. This dataset is currently accessible on the CMS website for the calendar years 2012 through 2015. The National Provider Identifier (NPI) is used to identify physicians [40], while Healthcare Common Procedure Coding System (HCPCS) codes are used to designate procedures. Other claims data comprises average payments and charges, the number of procedures performed, and medical specialty. CMS decided to aggregate Part B data based on: (1) the NPI of the performing provider, (2) the HCPCS code for the procedure or service conducted, and (3) the place of service, which is either a facility (F) or non-facility (O), such as a hospital or office, respectively. Each row in the dataset contains a physician's NPI, provider type, and one HCPCS code broken down by location of service, as well as information corresponding to this breakdown (i.e., claim counts) and other non-changing attributes. In practice, physicians perform the same procedure (HCPCS code) at both a facility and their office, and some physicians practice under multiple provider types (specialties), including Internal Medicine and Cardiology.

Therefore, there are as many rows for each physician as there are unique combinations of NPI, Provider Type, HCPCS code, and site of service, and Part B data can be considered to provide information at the procedure level. Table 1 provides an example of a physician from the 2015 Part B dataset with the NPI 1649387770.

Table 1 Sample of the Part B dataset

Npi	...	Provider_type	...	Place_of _service	Hcpcs _code	...	Line_srvc _cnt	...	Average_ submitted _chrg_amt	...
1649387770	...	Ophthalmology	...	O	66821	...	28	...	1200	...
1649387770	...	Ophthalmology	...	F	66984	...	154	...	2400	...
1649387770	...	Ophthalmology	...	O	67820	...	45	...	105	...
1649387770	...	Ophthalmology	...	O	76514	...	11	...	80	...
1649387770	...	Ophthalmology	...	O	92004	...	205	...	175	...

Part D

The Part D dataset contains information on the prescription drugs administered under the Medicare Part D Prescription Drug Program during a given year.

This data is currently available on the CMS website for the calendar years 2013 through 2015 (with 2015 being released in 2017) [47]. In the data, physicians are identified by their unique NPI, while drugs are identified by their brand and generic names. Other information includes average payments and charges, as well as variables describing the prescribed drug quantity and medical specialty. CMS has decided to aggregate the Part D data across (1) the prescriber's NPI and (2) the drug name (brand name in the case of trademarked drugs) and generic name. Each row of the Part D dataset contains a physician's NPI, provider type, and drug name, as well as information corresponding to this breakdown (i.e. claim counts) and other static attributes. Same as with Part B, we found a few physicians that practise under multiple specialties, such as Internal Medicine and Cardiology. Therefore, there are as many rows for each physician as there are unique combinations of NPI, Provider Type, drug name, and generic name, and Part D data can be considered to provide information at the procedure level. To safeguard the privacy of Medicare recipients, aggregated records derived from 10 or fewer claims are excluded from the Part

D data. Table 2 provides an example of a physician from the 2015 Part D dataset with the NPI 1649387770.

Table 2 Sample of Part D dataset

Npi	...	Provider_type	...	Drug_name	Total_drug _cost	Total_claim _count_ ge65	Ge65 _suppress _flag	...
1649387770	...	Ophthalmology	...	ALPHAGAN P	11811.27	57	NA	...
1649387770	...	Ophthalmology	...	AZASITE	3410.56	25	NA	...
1649387770	...	Ophthalmology	...	AZOPT	8336.27	27	NA	...
1649387770	...	Ophthalmology	...	BRIMONIDINE TAR TRATE	1769.25	12	NA	...
1649387770	...	Ophthalmology	...	COMBIGAN	25434.18	127	NA	...

DMEPOS

The DMEPOS dataset contains claims information regarding Medical Equipment, Prosthetics, Orthotics, and Supplies that physicians referred patients to either purchase or receive as a gift.

or rent from a vendor within a specified year. This dataset is based on claims submitted by suppliers to Medicare, whereas the physician's role is to refer the patient to the supplier. Currently, this information is accessible on the CMS website for the calendar years 2013 through 2015. (with 2015 being released in 2017) [48]. In the data, physicians are identified by their unique NPI, while products are designated by their HCPCS code. Other claims data includes average payments and charges, the number of services/products rented or sold, and medical specialty (also referred to as provider type). CMS chose to aggregate Part B data over: (1) the NPI of the performing provider, (2) the HCPCS code for the procedure or service performed by the DMEPOS supplier, and (3) the supplier rental indicator (value of 'Y' or 'N') derived from DMEPOS supplier claims (according to CMS documentation). Each row contains a physician's NPI, provider type, one HCPCS code divided by rental or non-rental, and specific information corresponding to this breakdown (e.g., the number of supplier claims) in addition to other non-changing attributes (i.e. gender). We have discovered that some physicians refer to the same DMEPOS equipment, or HCPCS code, as both rental and non-rental, as do a few physicians who practise under multiple specialties, such as Internal Medicine and Cardiology. Therefore, there are as many entries for each physician as there are unique combinations of NPI, Provider Type, HCPCS code, and rental status,

and the DMEPOS data can also be considered to provide information at the procedure level.

The physician whose NPI is 1649387770 is illustrated in Table 3 from the 2015 DMEPOS dataset.

Table 3 Sample of DMEPOS

Referring_npi	...	Referring _provider_ type	...	Hcpcs_code	...	Supplier _rental _indicator	Number_of _supplier _claims	Avg_supplier _submitted _charge	...
1649387770	...	Ophthalmol-ogy	...	V2020	...	N	44	67.4	...
1649387770	...	Ophthalmol-ogy	...	V2203	...	N	21	66.0	...
1649387770	...	Ophthalmol-ogy	...	V2303	...	N	18	87.5	...

LEIE

To accurately assess fraud detection performance in real-world practise, we need a data source that includes physicians who have committed real-world fraud.

fraud. Therefore, we utilise the List of Excluded Individuals and Entities (LEIE), which contains the following information: the reason for exclusion, the date of exclusion, and the reinstatement/waiver date for all current physicians deemed unfit to practise medicine and therefore excluded from practising in the United States for a specified time period.

In accordance with Sections 1128 and 1156 of the Social Security Act, this dataset was created and is updated monthly by the Office of Inspector General (OIG). The OIG has the authority to exclude individuals and organisations from federally-funded healthcare programmes like Medicare. Regrettably, the LEIE is not exhaustive, as 38 percent of providers with fraud convictions continue to practise medicine and 21 percent were not suspended despite their convictions. In addition, the LEIE dataset contains NPI values for a minority of physicians and entities. An example of four distinct physicians and how they are portrayed within the LEIE is shown in Table 4, where any physician without a listed \sNPI has a value of 0.

The LEIE is aggregated at the provider level and does not contain specific information regarding fraudulent procedures, medications, or equipment. There are various categories of exclusions, based on the severity of the offence, as described by distinct rule numbers. We do not utilise all exclusions, but rather filter excluded providers based on a subset of fraud-indicating criteria. The codes corresponding to fraudulent provider exclusions and the length of mandatory exclusion are listed in Table 5. We have determined that any conduct before and during the "end of exclusion date" of a physician constitutes fraud.

Table 4 Sample of LEIE

Specialty	...	Npi	...	Excltype	Excldate	...
GENERAL PRACTICE/FP	...	0	...	1128b6	19770701	...
EMPLOYEE	...	0	...	1128b6	19780124	...
GENERAL PRACTICE	...	1003016742	...	1128a1	20170720	...
NURSE/NURSES AIDE	...	1003011644	...	1128b4	20091220	...

Table 5 LEIE rules involving fraud

Rule number	Description
1128(a)(1)	Conviction of program-related crimes
1128(a)(2)	Conviction relating to patient abuse or neglect
1128(a)(3)	Felony conviction relating to health care fraud
1128(b)(4)	License revocation or suspension
1128(b)(7)	Fraud, kickbacks and other prohibited activities
1128(c)(3)(g)(i)	Conviction of two mandatory exclusion offenses 10 years
1128(c)(3)(g)(ii)	Conviction of 3 mandatory exclusion offenses indefinite

DATA PROCESSING

Part B was accessible between 2012 and 2015, whereas Part D and DMEPOS were accessible between 2013 and 2015. For Part B and DMEPOS, the initial phase consisted of removing all attributes absent from each available year.

In all available years, the Part D dataset contained the same attributes. As they were unavailable for other years, we removed the standard deviation variables from 2012 and 2013 and the standardised payment variables from 2014 and 2015. We removed a standard deviation variable for DMEPOS from 2014 and 2015 because it was unavailable in 2013. For each of the three datasets, we eliminated all instances that lacked both NPI and HCPCS/drug name values or contained an invalid NPI (i.e. NPI = 0000000000). For Part B, we removed all instances containing HCPCS codes for prescriptions. These prescription-related codes are for specific services listed in the Medicare Part B Drug Average Sales Price file, not actual medical procedures. Keeping these instances would muddy the results, as the line srvc_cnt feature in these cases represents the weight or volume of a substance, as opposed to counting the number of procedures.

In order to provide a solid framework for our experiments and analyses, we are only interested in specific attributes from each dataset for this study. For the Part B dataset, we retained eight features and eliminated the remaining twenty-two. Seven were retained for the Part D dataset, while the remaining fourteen were eliminated. For the DMEPOS dataset, we retained nine and eliminated the remaining nineteen. The excluded attributes include provider-related information, such as location and name, as well as redundant variables, such as text descriptions, that can be represented by the variables comprising the procedure or drug codes. For Part D, we also omitted variables that provided count and payment information for patients aged 65 or older, as this data is contained in the variables that were retained.

In this instance, the claim count variable (total claim count) includes estimates for patients aged 65 and older. Tables 6, 7, and 8 contain a description and feature type (numerical or categorical), as well as the exclusion attribute (fraud label) derived from the LEIE, for the features selected from the datasets.

The data processing stages for Part B, Part D, and DMEPOS are similar. All three unmodified datasets were aggregated by NPI and HCPCS/drug and originated at the HCPCS or procedure level.

Table 6 Description of features chosen from the Part B dataset

Feature	Description	Type
Npi	Unique provider identification number	Categorical
Provider_type	Medical provider's specialty (or practice)	Categorical
Nppes_provider_gender	Provider's gender	Categorical
Line_srvc_cnt	Number of procedures/services the provider performed	Numerical
Bene_unique_cnt	Number of distinct Medicare beneficiaries receiving the service	Numerical
Bene_day_srvc_cnt	Number of distinct Medicare beneficiary/per day services	Numerical
Average_submitted_chrg_amt	Average of the charges that the provider submitted for the service	Numerical
Average_medicare_payment_amt	Average payment made to a provider per claim for the service	Numerical
Exclusion	Fraud labels from the LEIE dataset	Categorical

Table 7 Description of features chosen from the Part D dataset

Feature	Description	Type
Npi	Unique provider identification number	Categorical
Specialty_description	Medical provider's specialty (or practice)	Categorical
Bene_count	Number of distinct Medicare beneficiaries receiving the drug	Numerical
Total_claim_count	Number of drug the provider administered	Numerical
Total_30_day_fill_count	Number of standardized 30-day fills	Numerical
Total_day_supply	Number of day's supply	Numerical
Total_drug_cost	Cost paid for all associated claims	Numerical
Exclusion	Fraud labels from the LEIE dataset	Categorical

Table 8 Description of features chosen from the DMEPOS dataset

Feature	Description	Type
Referring_npi	Unique provider identification number	Categorical
Referring_provider_type	Medical provider's specialty (or practice)	Categorical
Referring_provider_gender	Provider's gender	Categorical
Number_of_suppliers	Number of suppliers used by provider	Numerical
Number_of_supplier_beneficiaries	Number of beneficiaries associated by the supplier	Numerical
Number_of_supplier_claims	Number of claims submitted by a supplier from a referring order	Numerical
Number_of_supplier_services	Number of services/products rendered by a supplier	Numerical
Avg_supplier_submitted_charge	Average payment submitted by a supplier	Numerical
Avg_supplier_medicare_pmt_amt	Average payment awarded to suppliers	Numerical
Exclusion	Fraud labels from the LEIE dataset	Categorical

LEIE, we reorient each dataset, aggregating to the provider-level where all information is grouped by and aggregated over each NPI (and other specific features). For Part B, the aggregating process consists of grouping the data by NPI, provider type, gender and year, aggregating over HCPCS and place of service. Part D was grouped by NPI, provider type and year aggregating over drugs. DMEPOS was grouped by NPI, provider type, gender and year, aggregating over HCPCS and rental status. For the Part D and DMEPOS datasets, their beneficiary counts are suppressed to 0 if originally below 11, and in response we imputed the value of 5 as recommended by CMS.

In an effort to bypass information loss due to aggregating these datasets, we generated six numeric features for each chosen numeric feature outlined in the previous subsection for each dataset ("Medicare dataset descriptions" section). Therefore, for each numeric value, per year, in each dataset, we replace the original numeric variables with the aggregated mean, sum, median, standard deviation, minimum and maximum values, creating six new features for each original numeric feature. The resulting features are all complete except for standard deviation which contains NA values. These NA values are generated when a physician has performed/prescribed a HCPCS/drug once in a given year. Therefore, the population standard deviation for one unique instance is 0, and thus we replace all NA values with 0 representing that this single instance has no variability in that particular year. Two other features included are the categorical features: provider type and gender (Part D do not contain a gender variable).

Combined dataset

After processing Part B, Part D, and the DMEPOS datasets, the Combined dataset is generated, containing all the attributes from each, as well as the fraud labels derived from each.

under the LEIE. The process of combining entails a join on the NPI, provider type, and year. Due to the absence of a gender variable in the Part D data, we did not include this variable in the join operation conditions and instead utilised the gender labels from Part B, removing the gender labels from the DMEPOS dataset after the join. In combining these datasets, only physicians who have participated in all three Medicare parts can be considered. Our study demonstrates that this Combined dataset has a larger and more inclusive base of attributes for applying data mining algorithms to detect fraudulent behaviour.

Fraud labelling

We use the LEIE dataset to generate fraud labels for all four datasets. Only physicians within the LEIE dataset are considered fraudulent; otherwise, they are deemed nonfraudulent.

To obtain exact matches between the Medicare datasets and the LEIE, we determined that the NPI value is the only way to precisely match physicians, thereby ensuring the highest level of data integrity. The LEIE provides specific dates (month/day/year) for when the exclusion begins and the duration of the exclusion period, whereas we use only month/year (no rounding within a month; for example, May 1st through May 31st is considered May). For instance, if a provider violates rule 1128(a)(3) ('felony conviction due to healthcare fraud'), which carries a minimum exclusion period of five years beginning in February 2010, the exclusion period would end in February 2015. Note that we utilised the earliest date between the exclusion end date (based on the minimum exclusion period added to the date of initiation), the waiver date, and the reinstatement date.

Consequently, if a waiver date of October 2014 and a reinstatement date of December 2014 are also listed, the exclusion period would be between February 2010 and October 2014. This accounts for providers who may still be in their exclusion period but have received a waiver or reinstatement to use Medicare and are therefore no longer considered fraudulent as of this waiver or reinstatement date or later.

Contrary to the LEIE data, the Medicare datasets are released annually where all data is provided for each given year. In order to best handle the disparity between the annual and monthly dates, we round the new exclusion end date to the nearest year based on the month. If the end exclusion month is greater than 6 (majority of the year), then the exclusion end year is increased to the following year; otherwise, the current year is used. We do not want a physician to be considered fraudulent during a year unless more than half that year is before their exclusion end date. Continuing the above example, we determined that the end exclusion date was October 2014, therefore since October is the tenth month and 10 is greater than 6, the end exclusion year would be rounded up to 2015. Therefore, translating this to the Medicare data, any activity in 2014 or earlier would be considered fraudulent when creating fraud labels. For further clarification, if the waiver date would have been March 2014, the

end exclusion year would be 2014 and only activity from 2013 or earlier would be labelled fraudulent.

Table 9 Distribution of fraud labels

Dataset	Non-fraudulent	Fraudulent	% Fraudulent
Part B	3,691,146	1409	0.038
Part D	2,098,715	1018	0.048
DMEPOS	862,792	635	0.074
Combined	759,267	473	0.062

The LEIE dataset is joined to all four datasets based on NPI. We create an exclusion feature which is the final categorical attribute discussed in previous sections, which indicates either fraud or non-fraud instances. Any physician practicing within a year prior to their exclusion end year is labelled fraudulent.

One-hot encoding

In order to build our models with a combination of numerical and categorical features, we employ one-hot encoding, transforming the categorical features. For example, one-hot encoding gender would first consist of generating extra features equaling the number of options, in this case two (male and female). If the physician is male, the new male feature would be assigned a 1 and the female feature would be 0; while for female, the male would be assigned a 0 and the female assigned a 1. If the original gender feature is missing then both male and female are assigned a 0. This process is done for all four datasets for gender and provider type/specialty. Table 10 summarizes all four datasets after data processing and after the categorical features have been one-hot encoded. Note that NPI is not used for building models and is removed from each dataset after this step.

Table 10 Summary of Medicare datasets

	Part B		Part D		DMEPOS		Combined	
	Instances	Features	Instances	Features	Instances	Features	Instances	Features
After processing and fraud labeling	3,692,555	35	2,099,733	34	863,427	41	759,740	102
After one-hot encoding	3,692,555	126	2,099,733	126	863,427	145	759,740	173

Methodology

Learners

Due to the Huge Volume of the datasets, we used Spark on top of a Hadoop Yarn cluster for running and validating models. We used the three available classification models.

Logistic Regression, Gradient Boosted Trees, and Random Forest are available in the Apache Spark 2.3.0 Machine Learning Library. In this section, we provide a concise description of each learner and highlight any configuration changes that deviate from the default settings.

Logistic Regression (LR) predicts probabilities for which class a categorical dependent variable belongs to by employing a logistic function and a set of independent variables. LR generates values that can be interpreted as class probabilities by utilising a sigmoidal (logistic) function. LR is comparable to linear regression, but it utilises a distinct hypothesis class to predict class membership. The bound matrix was configured to match the geometry of the data (number of classes and features), so the algorithm is aware of the number of classes and features contained in the dataset. For binomial regression, the bound vector size is equal to 1, and no thresholds are set for binary classification.

Random Forest (RF) is a method of ensemble learning that produces a significant number of trees. The class value that appears most often among these trees is the class predicted as the model's output. As a method for ensemble learning, RF is a collection of numerous tree predictors. Each tree in the forest is dependent on the values determined by an independently sampled random vector, and each tree is equally distributed throughout the forest. Randomness is introduced into the training procedure by the RF ensemble, which can reduce overfitting and is reasonably robust to imbalanced data. Each RF learner is constructed with 100 trees, as our research has shown that adding more trees provides little to no benefit. To reduce training time, the parameter that caches node IDs for each instance was set to true, and the maximum memory parameter was set to 1024 MB. The setting that controls the number of features to consider for splits at each tree node was adjusted to one-third, as this setting yielded superior initial results. The maximum bins parameter, which specifies the maximum number of bins for discretizing continuous features, is set to 2 because categorical features have been converted using onehot encoding and are no longer present.

Gradient Boosted Trees (GBT) is a decision tree ensemble. In contrast to RF, GBT trains each decision tree independently in order to minimise loss as determined by the loss function of the algorithm. During each iteration, the current ensemble is used to predict the class for each instance in the training data. The predicted values are evaluated \switch the actual values allowing the algorithm to pinpoint and correct previously \semilabeled instances. The parameter that caches node IDs for each instance, was set to \strue and the maximum memory parameter was set to 1024 MB to minimise training \stime.

Performance metric

In assessing Medicare fraud, we are presented with a two-class classification problem where a physician is either fraudulent or non-fraudulent. In our study, the positive class, or class of interest, is fraud and the negative class is non-fraud. Spark presented us with a confusion matrix for each model and is commonly used to assess the performance of learners. Confusion matrices provide counts comparing actual counts against predicted counts. From the resultant matrices, we employ AUC to measure fraud detection performance. AUC is the Area under the Receiver Operating Characteristic (ROC) curve, where ROC is the comparison between false positive (fall-out) and true positive (recall). Recall is calculated by TP
TP+FN and fall-out is calculated by FP
FP+TN. The definitions
for TP, TN, FP and FN, which can be directly calculated from the confusion matrix are as follows:

•• True positive (TP): number of actual positive instances correctly predicted as positive.
•• True negative (TN): number of actual negative instances correctly predicted as negative.
•• False positive (FP): number of negative instances incorrectly classified as positive.
•• False negative (FN): number of positive instances incorrectly assigned as negative.

The AUC curve is an encompassing evaluation of a learner as it depicts performance across all decision thresholds. The AUC results in a single value ranging from 0 to 1, where a perfect classifier results in an AUC of 1, an AUC of 0.5 is equivalent to random guessing and less than 0.5 demonstrates bias towards a given class. AUC has been found to be effective for class imbalance.

Cross-validation

We employ stratified k-fold cross-validation in evaluating our models, where k = 5. Stratification ensures all folds have class representation matching the ratio of the original data, which is important when dealing with largely imbalanced data. The training data is evenly divided into fivefold where fourfold will be used for training the model and the remaining fold tests the model. This process is repeated 5 times allowing each fold an opportunity as the test fold, ensuring the entire dataset is fully leveraged being used in training and validation. Spark will automatically create different folds each time the learner is run, and to validate our results we ran each model 10 times for each learner/ dataset pair. The use of repeats helps to reduce bias due to bad random draws when creating the folds where the final performance for every presented result is the average over all 10 repeats.

Results and dialogue

This section discusses the findings of our study evaluating Medicare fraud detection dataset and learner performance. Individual physicians' practises are unique.

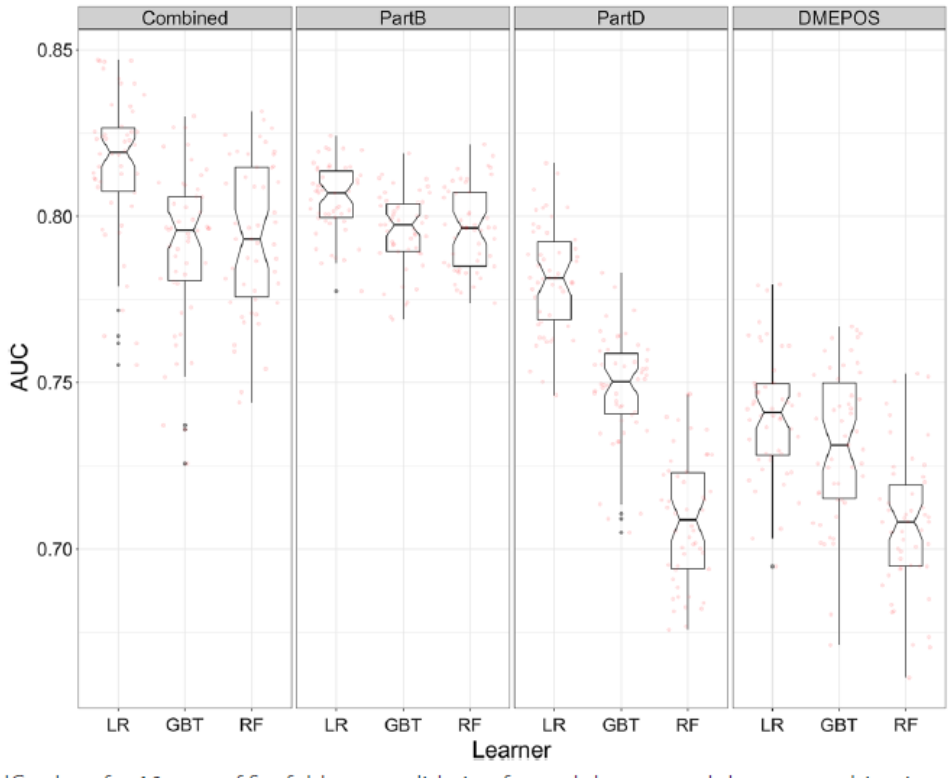
where a physician may submit claims to Medicare only via Part B, Part D, or DMEPOS, or all three. In order to determine the optimal combinations for detecting Medicare fraud, we compare learner performance to each Medicare dataset. The AUC results for each dataset and learner combination are displayed in Table 11. The values in italics represent the highest AUC scores per dataset, while the values in bold represent the highest per learner. LR produces the two highest overall AUC scores for the Combined and Part B datasets with 0.816 and 0.805, respectively. The Combined dataset has the highest AUC overall, but the Part B dataset demonstrates the least variation in fraud detection performance across learners, including the highest AUC scores for GBT and RF. The Part D and DMEPOS datasets have the lowest AUC values for all three learners, but LR and GBT perform better than RF when analysing these datasets.

Table 11 Learner AUC results by dataset

Dataset	Logistic Regression	Gradient Boosted Trees	Random Forest
Combined	*0.81554*	0.79047	0.79383
Part B	*0.80516*	0.79569	0.79604
Part D	*0.78164*	0.74851	0.70888
DMEPOS	*0.74063*	0.73129	0.70756

The favourable results obtained using LR with each of the datasets may be attributable to the squarederror loss function with the application of L2 regularisation, also known as Ridge Regression, which penalises large coefficients and improves generalisation performance, thereby rendering LR fairly robust to noise and overfitting. Even though LR performs well on the Part B and Combined datasets, additional testing is necessary to determine if the Part D and DMEPOS datasets have unique characteristics that contribute to their inferior fraud detection performance. The poor performance of tree-based methods, especially RF, may be attributable to the lack of independence between individual trees or the large number of categorical variables. The Combined dataset contains features from all three parts of Medicare, resulting in a robust pool of attributes and presumably enhancing model generalisation and fraud detection performance. In particular, the Combined dataset utilising LR has the highest AUC and superior performance compared to each of its Medicare components. Part B displays the greatest AUC scores for RF and GBT. Intriguingly, the Part B dataset has the lowest inter- and intra-learner variability, which may be partially attributable to having the highest number of fraud classifications. Not only do the Part D and DMEPOS datasets demonstrate weak learner performance, but the AUC variability across learners is generally higher. This may indicate potential negative effects of a high-class imbalance or less discriminatory power in the chosen characteristics. Figure below depicts a

box plot of our experimental results for all 50 AUC values from the ten trials of fivefold cross-validation for each dataset/learner pair.



The results of the two-factor ANOVA test for each Dataset and Learner, as well as their interaction, are shown in Table 12. (Dataset: Learner). The ANOVA test demonstrates, with a 95% confidence interval, that these factors and their interactions are statistically significant. To determine statistical groupings, we apply the Tukey's HSD test to the Medicare datasets' results, which confirms the superior performance of the LR learner and the Combined dataset for Medicare fraud detection (as seen in Table 11).

Table 12 Two-factor ANOVA test results

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Dataset	3	0.6257	0.20855	594.15	< 2e−16
Learner	2	0.1174	0.05868	167.17	< 2e−16
Dataset:Learner	6	0.0658	0.01097	31.26	< 2e−16
Residuals	588	0.2064	0.00035	–	–

Table 13 Two-factor Tukey's HSD learner results over all datasets

Learner	Group	AUC	sd	r	Min	Max
Logistic Regression	a	0.78574	0.03369	200	0.69487	0.847
Gradient Boosted Trees	b	0.76649	0.03343	200	0.67119	0.83013
Random Forest	c	0.75158	0.04753	200	0.66138	0.83161

Table 13 demonstrates that LR is substantially superior to GBT and RF for each learner across all datasets. LR and GBT have comparable AUC variability, but LR has the highest minimum and maximum AUC scores, further demonstrating the superior performance of LR for each dataset. Table 14 summarises the significance of each student's dataset performance. The Combined and Part B datasets perform significantly better than the Part D and DMEPOS datasets, while the DMEPOS dataset performs significantly worse than the Part D dataset. Since the Part B and Combined results do not differ significantly, we prefer the Combined dataset for general fraud detection, as it is difficult to predict in advance which part of the Medicare system a physician or provider will target for fraudulent activity (e.g., medical procedures/services, drug submissions, or prosthetic rental).

With the Combined dataset, we have a larger web for monitoring fraudulent behaviour, as opposed to monitoring just one Medicare component for a given healthcare provider.

Table 14 Two-factor Tukey's HSD dataset results over all learners

Dataset	Group	AUC	sd	r	Min	Max
Combined	a	0.79995	0.02549	150	0.7258	0.847
Part B	a	0.79896	0.0123	150	0.769	0.82425
Part D	b	0.74634	0.03443	150	0.67576	0.81602
DMEPOS	c	0.72649	0.02506	150	0.66138	0.77957

In addition, the Combined dataset with LR yields the only results in which the Combined dataset achieves superior performance to the individual Medicare datasets. Consequently, based on these exploratory performance results, we demonstrate that when a physician has participated in Part B, Part D, and DMEPOS, the Combined dataset employing LR provides the greatest overall fraud detection performance.

Conclusion

Since the number of people aged 65 and older in the United States continues to rise, there is an increasingly urgent need to cut down on Medicare fraud overall. This is of the utmost importance in the United States. Because Medicare is essential for a large number of citizens, there is a significant emphasis placed on high-quality research.

into fraud detection, with the goal of maintaining fair and acceptable rates for healthcare. Throughout the course of an ever-increasing number of years, the Centers for Medicare & Medicaid Services (CMS) has made many Big Data Medicare claims datasets available for public use. Throughout this body of work, we present a novel approach (combining multiple Medicare datasets and leveraging state-of-the-art Big Data processing and machine learning approaches) for determining the fraud detection capabilities of three Medicare datasets, individually and combined, utilising three learners, against real-world fraudulent physicians and other medical providers taken from the LEIE dataset. These capabilities are tested against real-world fraudulent physicians and other medical providers.

We discuss the algorithms that we have developed for analysing each dataset provided by CMS, as well as the Combined dataset, and the mapping of provider fraud labels. Experiments were conducted on all four datasets, Combined, Part B, Part D, and DMEPOS respectively. Because each dataset qualified as Big Data, we were forced to execute and validate our models using Spark atop a Hadoop YARN cluster. This was necessary in order to process the data. Each dataset was trained and evaluated using three different learners: logistic regression, gradient-boosted trees, and random forests. The Combined dataset had the best overall fraud detection performance with an AUC of 0.816 using LR, indicating better performance than each of its individual Medicare parts, and scored similarly to Part B with no significant difference in average AUC. This indicates that the Combined dataset had better performance than each of its individual Medicare parts. The DMEPOS dataset got the worst overall outcomes for all of the students in the study. Consequently, as a result of these experimental findings and observations, in conjunction with the idea that a physician or provider might commit fraud via any element of Medicare, we show that utilising the Combined dataset with LR delivers the best overall fraud detection performance. In upcoming work, we will be applying data sampling approaches in order to address the uneven nature of known fraud incidents while reviewing the various Medicare datasets. This will be done in an effort to improve accuracy.

Future works

There are several possible future works that can be done on Medicare fraud detection, including:

Improving Machine Learning Models: One area of focus could be on developing more advanced machine learning algorithms that can better detect patterns of fraud in Medicare claims data. This could involve exploring new techniques, such as deep learning, natural language processing, and graph analysis.

Incorporating Unstructured Data: Another possible area of improvement is to incorporate unstructured data sources such as social media, news articles, and other publicly available data to improve fraud detection algorithms.

Real-time Fraud Detection: Developing real-time fraud detection systems that can monitor transactions as they occur and flag suspicious activity immediately could be another area of focus. This could involve the use of real-time data processing technologies such as stream processing.

Collaboration between Private and Public Sectors: Collaboration between the private and public sectors could help to improve Medicare fraud detection. Private companies that have developed fraud detection technologies could share their tools and expertise with government agencies responsible for monitoring Medicare fraud.

User-Friendly Reporting Tools: Providing user-friendly tools for healthcare providers to report suspected fraud can encourage more reporting and improve the overall effectiveness of fraud detection efforts.

Applying Blockchain technology: Using blockchain technology to secure Medicare's claim data can make it more difficult for fraudsters to manipulate or create fraudulent claims.

AI-Enabled Monitoring: Integrating AI-enabled monitoring tools that can constantly analyse Medicare claims for irregularities can help detect fraud more efficiently.

These are just a few of the possible future works that could be done to improve Medicare fraud detection. The field is constantly evolving, and new techniques and technologies are emerging all the time, so there is a lot of potential for innovation in this area.

REFERENCES

1) https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0138-3
2) https://ieeexplore.ieee.org/document/8848229
3) https://www.semanticscholar.org/paper/Healthcare-Insurance-Fraud-Detection-Leveraging-Big-Dora-Sekharan/88458ea65aadf372e9f6856d65b29c3c9509e0c4
4) https://paperswithcode.com/paper/bigdl-a-distributed-deep-learning-framework
5) CMS: Research, Statistics, Data, and Systems. http://www.cms.gov/resea rch-stati stics -data-and-syste ms/resea rchstati stics -data-and-syste ms.html.
6) Medicare.gov. What's medicare. https ://www.medic are.gov/sign-up-chang e-plans /decid e-how-to-get-medic are/whats -medic are/what-is-medic are.html.
7) Centers for Medicare & Medicaid Services. Medicare fraud & abuse: prevention, detection, and reporting. https ://www.cms.gov/Outre ach-and-Educa tion/Medic are-Learn ing-Netwo rk-MLN/MLNPr oduct s/downl oads/fraud_and_abuse .pdf
8) LEIE: Office of Inspector General LEIE Downloadable Databases. https ://oig.hhs.gov/exclu sions /autho ritie s.asp.